UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
10/033,303	10/19/2001	Luke V. Schneider	020444-001310US	8489

20350 7590 05/25/2006

TOWNSEND AND TOWNSEND AND CREW, LLP
TWO EMBARCADERO CENTER
EIGHTH FLOOR
SAN FRANCISCO, CA 94111-3834

EXAMINER	
SISSON, BRADLEY L	

ART UNIT	PAPER NUMBER
1634	

DATE MAILED: 05/25/2006

Please find below and/or attached an Office communication concerning this application or proceeding.

Interview Summary	Application No.	Applicant(s)	
	10/033,303	SCHNEIDER ET AL.	
	Examiner	Art Unit	
	Bradley L. Sisson	1634	

All participants (applicant, applicant's representative, PTO personnel):

(1) *Bradley L. Sisson*. (3)*Luke V. Schneider, Ph.D.*.

(2) *Kenneth E. Jenkins, Ph.D., Reg. No. 51,846*. (4)_____.

Date of Interview: *16 May 2006*.

Type: a)☒ Telephonic b)☐ Video Conference
c)☐ Personal [copy given to: 1)☐ applicant 2)☐ applicant's representative]

Exhibit shown or demonstration conducted: d)☒ Yes e)☐ No.
If Yes, brief description: *Powerpoint presentation "Target Discovery from OMICS to Knowmics," 7 frames, printout attached*.

Claim(s) discussed: *1-3,5-9,13-21,25,26,67-96,104 and 107-128*.

Identification of prior art discussed: *"Algorithms for de novo peptide sequencing via tandem mass spectrometry," Lu and Chen. Provided via email 17 May 2006*.

Agreement with respect to the claims f)☐ was reached. g)☒ was not reached. h)☐ N/A.

Substance of Interview including description of the general nature of what was agreed to if an agreement was reached, or any other comments: *See Continuation Sheet*.

(A fuller description, if necessary, and a copy of the amendments which the examiner agreed would render the claims allowable, if available, must be attached. Also, where no copy of the amendments that would render the claims allowable is available, a summary thereof must be attached.)

THE FORMAL WRITTEN REPLY TO THE LAST OFFICE ACTION MUST INCLUDE THE SUBSTANCE OF THE INTERVIEW. (See MPEP Section 713.04). If a reply to the last Office action has already been filed, APPLICANT IS GIVEN A NON-EXTENDABLE PERIOD OF THE LONGER OF ONE MONTH OR THIRTY DAYS FROM THIS INTERVIEW DATE, OR THE MAILING DATE OF THIS INTERVIEW SUMMARY FORM, WHICHEVER IS LATER, TO FILE A STATEMENT OF THE SUBSTANCE OF THE INTERVIEW. See Summary of Record of Interview requirements on reverse side or on attached sheet.

Examiner Note: You must sign this form unless it is an Attachment to a signed Office action.

B. L. Sisson
Examiner's signature, if required

Summary of Record of Interview Requirements

Manual of Patent Examining Procedure (MPEP), Section 713.04, Substance of Interview Must be Made of Record
A complete written statement as to the substance of any face-to-face, video conference, or telephone interview with regard to an application must be made of record in the application whether or not an agreement with the examiner was reached at the interview.

Title 37 Code of Federal Regulations (CFR) § 1.133 Interviews
Paragraph (b)
In every instance where reconsideration is requested in view of an interview with an examiner, a complete written statement of the reasons presented at the interview as warranting favorable action must be filed by the applicant. An interview does not remove the necessity for reply to Office action as specified in §§ 1.111, 1.135. (35 U.S.C. 132)

37 CFR §1.2 Business to be transacted in writing.
All business with the Patent or Trademark Office should be transacted in writing. The personal attendance of applicants or their attorneys or agents at the Patent and Trademark Office is unnecessary. The action of the Patent and Trademark Office will be based exclusively on the written record in the Office. No attention will be paid to any alleged oral promise, stipulation, or understanding in relation to which there is disagreement or doubt.

The action of the Patent and Trademark Office cannot be based exclusively on the written record in the Office if that record is itself incomplete through the failure to record the substance of interviews.

It is the responsibility of the applicant or the attorney or agent to make the substance of an interview of record in the application file, unless the examiner indicates he or she will do so. It is the examiner's responsibility to see that such a record is made and to correct material inaccuracies which bear directly on the question of patentability.

Examiners must complete an Interview Summary Form for each interview held where a matter of substance has been discussed during the interview by checking the appropriate boxes and filling in the blanks. Discussions regarding only procedural matters, directed solely to restriction requirements for which interview recordation is otherwise provided for in Section 812.01 of the Manual of Patent Examining Procedure, or pointing out typographical errors or unreadable script in Office actions or the like, are excluded from the interview recordation procedures below. Where the substance of an interview is completely recorded in an Examiners Amendment, no separate Interview Summary Record is required.

The Interview Summary Form shall be given an appropriate Paper No., placed in the right hand portion of the file, and listed on the "Contents" section of the file wrapper. In a personal interview, a duplicate of the Form is given to the applicant (or attorney or agent) at the conclusion of the interview. In the case of a telephone or video-conference interview, the copy is mailed to the applicant's correspondence address either with or prior to the next official communication. If additional correspondence from the examiner is not likely before an allowance or if other circumstances dictate, the Form should be mailed promptly after the interview rather than with the next official communication.

The Form provides for recordation of the following information:
- Application Number (Series Code and Serial Number)
- Name of applicant
- Name of examiner
- Date of interview
- Type of interview (telephonic, video-conference, or personal)
- Name of participant(s) (applicant, attorney or agent, examiner, other PTO personnel, etc.)
- An indication whether or not an exhibit was shown or a demonstration conducted
- An identification of the specific prior art discussed
- An indication whether an agreement was reached and if so, a description of the general nature of the agreement (may be by attachment of a copy of amendments or claims agreed as being allowable). Note: Agreement as to allowability is tentative and does not restrict further action by the examiner to the contrary.
- The signature of the examiner who conducted the interview (if Form is not an attachment to a signed Office action)

It is desirable that the examiner orally remind the applicant of his or her obligation to record the substance of the interview of each case. It should be noted, however, that the Interview Summary Form will not normally be considered a complete and proper recordation of the interview unless it includes, or is supplemented by the applicant or the examiner to include, all of the applicable items required below concerning the substance of the interview.

A complete and proper recordation of the substance of any interview should include at least the following applicable items:
1) A brief description of the nature of any exhibit shown or any demonstration conducted,
2) an identification of the claims discussed,
3) an identification of the specific prior art discussed,
4) an identification of the principal proposed amendments of a substantive nature discussed, unless these are already described on the Interview Summary Form completed by the Examiner,
5) a brief identification of the general thrust of the principal arguments presented to the examiner,
   (The identification of arguments need not be lengthy or elaborate. A verbatim or highly detailed description of the arguments is not required. The identification of the arguments is sufficient if the general nature or thrust of the principal arguments made to the examiner can be understood in the context of the application file. Of course, the applicant may desire to emphasize and fully describe those arguments which he or she feels were or might be persuasive to the examiner.)
6) a general indication of any other pertinent matters discussed, and
7) if appropriate, the general results or outcome of the interview unless already described in the Interview Summary Form completed by the examiner.

Examiners are expected to carefully review the applicant's record of the substance of an interview. If the record is not complete and accurate, the examiner will give the applicant an extendable one month time period to correct the record.

Examiner to Check for Accuracy

If the claims are allowable for other reasons of record, the examiner should send a letter setting forth the examiner's version of the statement attributed to him or her. If the record is complete and accurate, the examiner should place the indication, "Interview Record OK" on the paper recording the substance of the interview along with the date and the examiner's initials.

Continuation of Substance of Interview including description of the general nature of what was agreed to if an agreement was reached, or any other comments: Dr. Schneider provided a tutorial on the background of the invention and the application of the cumulative ranking algorithm. Dr. Schneider indicated that prior art methods would store the m/z values for the first 5 possible combinations of amino acids, but that the values become to voluminous and incapable of storage on a computer when one starts looking at the 6th position, or even further out.

Dr. Schneider stated how the values of the m/z charges observed/measured are not stored, but rather, the values are calculated anew for each fragment as it passes into the detection chamber of the mass spec. Dr. Schneider indicated that only the mean and standard deviation values are stored in the computer, and that the further one sequences, the greater the confidence interval one has as to the sequence of the preceding amino acid residues. However, the confidence interval is not necessarily all that great when one has reached the end of a 5-mer or 6-mer as there are no additional residues upon which to base m/z values.

Dr. Schneider indicated that in some instances you have the same, or nearly the same, m/z value for different fragments. The aspect of not being able to identify Leu/Ile was specifically discussed. Dr. Schneider indicated that they adapted a genetic algorithm to the current application.

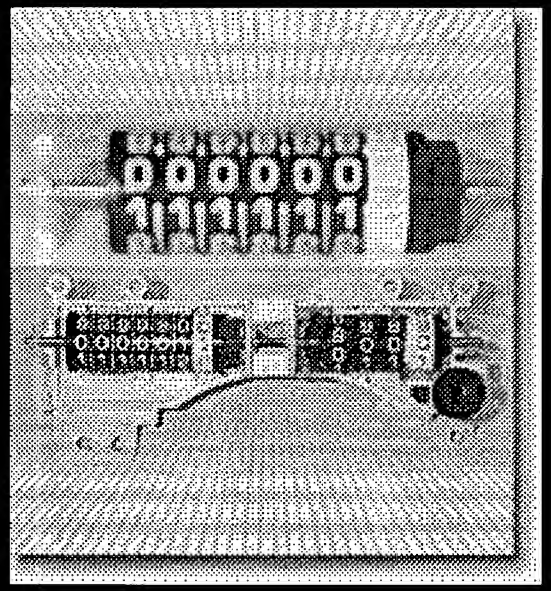
Dr. Schneider indicated that a mass spec is limited as to just how far it can see, noting that the mass spec used in the examples could only see out 6 amino acids. Dr. Schneider identified "Top Down Sequencing" as being developed post filing, and that employing this post filing technology, one can see out to 50 amino acid residues.

In response to inquiry by Mr. Sisson as to the ability to accurately sequence any polymer, and citing the 1-4, or 1-6 glycosidic linkages found in polysaccharides, Dr. Schneider stated that using mass spec, one cannot identify the linkage nor identify the saccharide beyond its being a hexose. Dr. Schneider added that post filing, they have gone to using PEG (polyethylene glycol) as a standard in MALDI mass spec.

Dr. Schneider indicated that sequencing is "probabilistic" and that the top ranked sequence may not be the correct sequence, directing attention to Example 13 of the specification. The aspect of not being able to identify the correct sequence when the target is an unknown was discussed. Dr. Schneider indicated that one could take the ranked sequencing values and then go search databases to se if something like it was known.

The Power of Cumulutive Ranking in Sequencing Oligomers

Exhaustive Search: Odometer Model



1 = A
2 = C
.
.
.
20 = Y

Lookup Counts for all sequences of length i

↓

Calculate Mean & σ

$$\text{Mean}_i = \frac{\sum \text{Counts}_i}{19^i}$$

$$\sigma = \sqrt{\frac{\left(\sum \text{Counts}_i^2\right) - \left(\sum \text{Counts}_i / 19^i\right)}{\left(19^i - 1\right)}}$$

↓

Lookup Counts for all sequences of length i

$$\frac{\text{Counts}_{i,j} - \text{Mean}_i}{\sigma_i} \Rightarrow p_{i,j}$$

Cum Rank Discriminates Alternatives

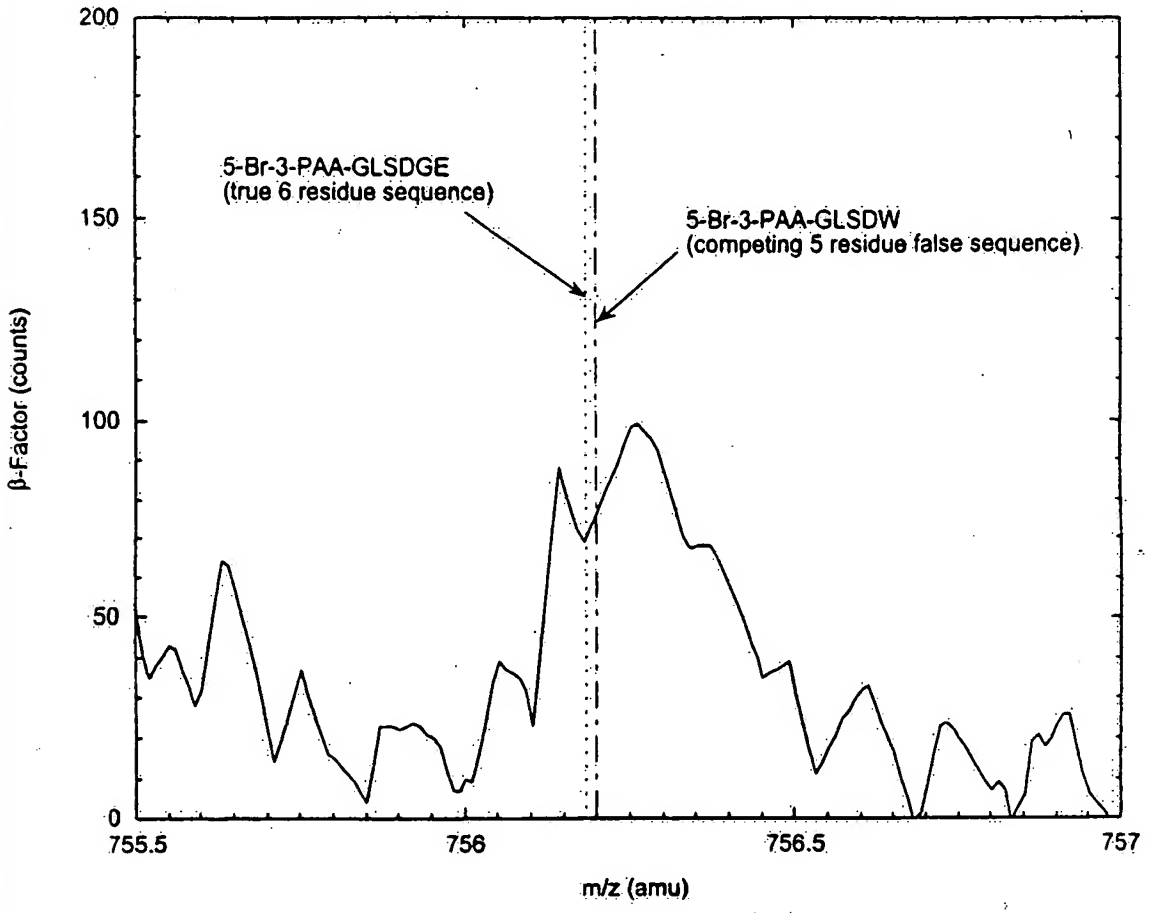
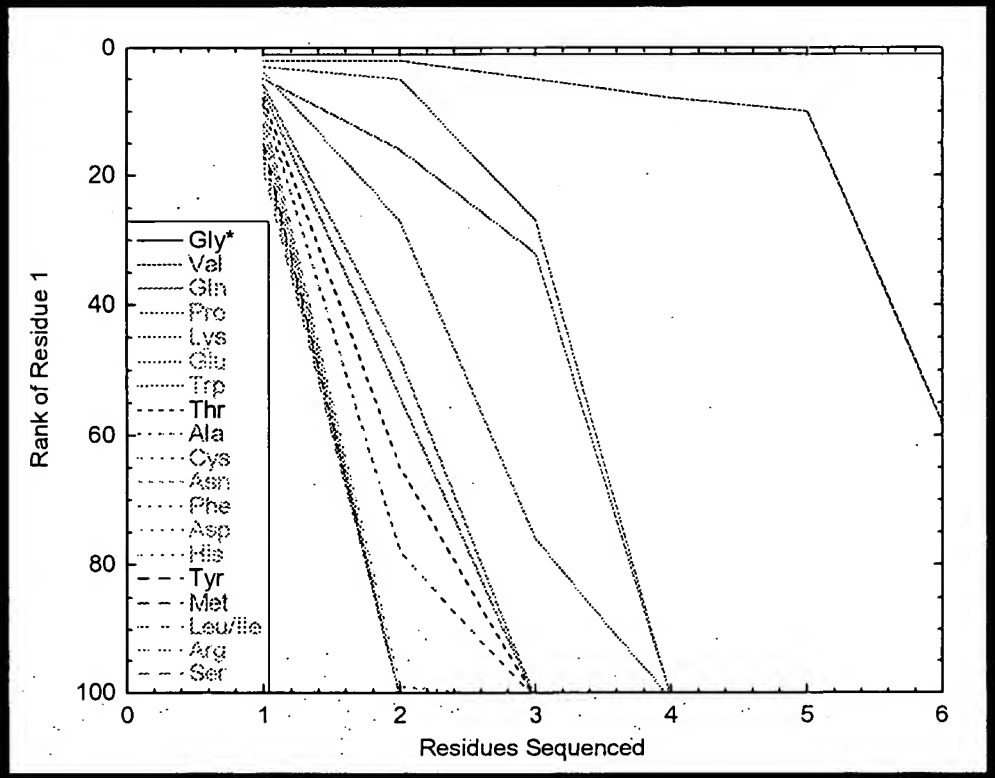


FIG. 33

Cumulative Ranking

- Cumulative ranking strengthens the identification of the earlier amino acids in the sequence by lowering the rank of competing sequences.
- Earlier amino acids are known to a higher certainty than later ones in the sequence as sequencing continues forward.

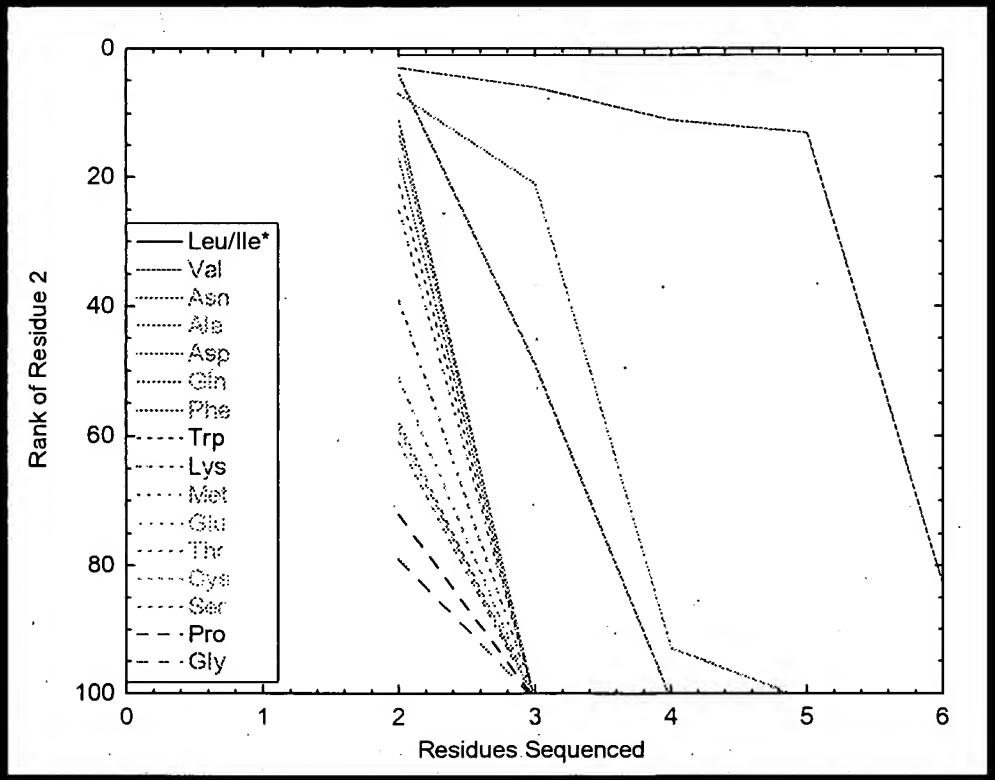
Rank of All Sequence Possibilities at Residue 1



BrPyr-labeled Myoglobin

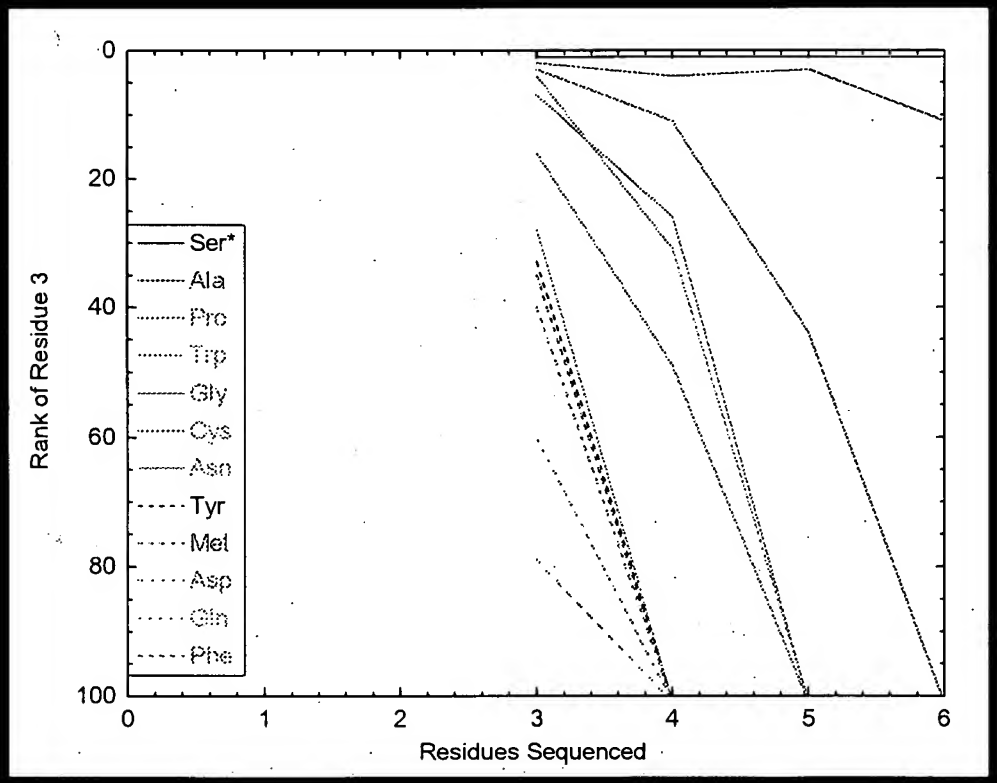
Residue 2

Rank of All Possibilities at Residue 2



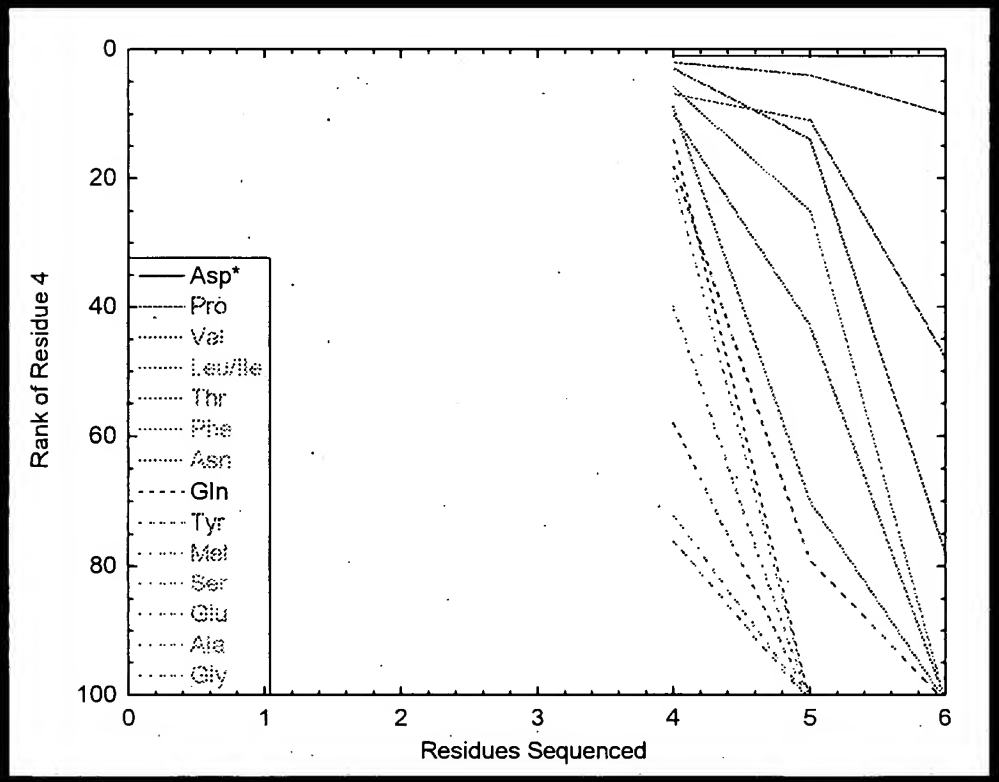
Residue 3

Rank of All Possibilities at Residue 3



Residue 4

Rank of All Possibilities at Residue 4



Sisson, Bradley

From: Jenkins, Kenneth E. [kejenkins@townsend.com]
Sent: Wednesday, May 17, 2006 12:57 PM
To: Sisson, Bradley
Subject: Review Article

<<DrugDiscovery-denovoreview-2004.pdf>>
Dear Examiner Sisson-

It was a pleasure talking with you yesterday. Per our meeting, enclosed is a review article which you may find helpful.

I will follow up with you on Monday.

Best regards,
Ken

Kenneth E. Jenkins, Ph.D.
Attorney at Law
Townsend and Townsend and Crew LLP
12730 High Bluff Drive Suite 400
San Diego, CA 92130
858.350.6100
kejenkins@townsend.com

Algorithms for *de novo* peptide sequencing via tandem mass spectrometry

Bingwen Lu Ting Chen *

Abstract

There is growing interest in the qualitative and quantitative analysis of proteins on a proteome-wide scale. Mass spectrometry plays an important role in the high-throughput environments of proteomics study. There have been two major developments for mass spectrometry technology: (1) instrumental (including physiochemical) development, such as the development of chromatography technology, the development of protein ionization technology, and the development of protein-labeling technology, and (2) the development of computational algorithms for analysis of mass spectra produced by mass spectrometers. In this review, we will give an overview of the development of computational algorithms for *de novo* peptide sequencing using tandem mass spectrometry.

*Corresponding author. Address: Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089 USA. Email: tingchen@hto.usc.edu. Tel: 213-740-2415.

Keywords: *de novo* peptide sequencing, mass spectrometry, proteomics, graph theory, dynamic programming, suboptimal algorithms

Teaser: Special focus on *de novo* peptide sequencing: a useful technique for proteomics study and peptide drug discovery

Mass spectrometry and proteomics

Mass spectrometry is an indispensable tool in the era of modern proteomics research. Mass spectrometry has been applied successfully to the study of proteins in the proteome-wide manner (*e.g.*, see [1-6]). To take one example, Gavin *et al.* [3] used tandem affinity purification (TAP) and mass spectrometry to characterize multi-protein complexes in *Saccharomyces cerevisiae*. They processed 1,739 genes and purified 589 protein assemblies. Their analysis of these assemblies revealed 232 distinct multi-protein complexes, and they proposed cellular roles for 344 proteins, among which 231 had no previous functional annotation. To take a second example, Ho *et al.* [4] employed a method called high-throughput mass spectrometric protein complex identification (HMS-PCI) to systematically identify protein complexes in *Saccharomyces cerevisiae*. Starting with 10% of the predicted proteins, they detected 3,617 associated proteins covering 25% of the yeast proteome. A third example is the analysis by Petricoin *et al.* [5] of the mass spectrometry-generated proteomic pattern in serum for ovarian cancer diagnosis. A training set of mass spectra generated from 50 unaffected women and 50 ovarian cancer patients were analyzed by an artificial-intelligence algorithm to discover a proteomic pattern that could completely discriminate cancer samples from normal samples. The identified pattern was then used to classify masked serum samples. The authors successfully

identified all 50 ovarian cancer samples, while for the 66 cases of non-ovarian cancer 63 were classified as non-cancer. They computed the positive predictive value for this validation to be 94%. This is quite comparable to another ovarian cancer diagnostic measure called CA125, which has a positive predictive value of 34%.

Protein sequencing and identification by tandem mass spectrometry

In a routine protein sequencing and identification by tandem mass spectrometry, the protein or proteins are first digested by some enzymes such as trypsin. The resulting peptides are then separated by liquid chromatography (LC) and subsequently analyzed by a mass spectrometer. The peptides are ionized, and the mass-to-charge (m/z) ratios are measured. For tandem mass spectrometry, ions within some range of specific m/z ratios are further selected, and a second round of fragmentation and m/z measuring is performed. The resulting tandem mass spectra, comprising m/z ratios and corresponding intensities, are then used for the identification of the original protein or peptide. There are two principal ways of doing this kind of identification. One is to correlate the spectra with protein sequences or nucleic acid sequences by database searching [7-12]. The other is to derive the peptide sequence directly, without the help of any sequence database. The latter method is called *de novo* peptide sequencing, which we will discuss extensively below.

The *de novo* peptide sequencing problem and the scoring function

The *de novo* peptide sequencing problem, given a tandem mass spectrum, is that of seeking a peptide that can best explain the spectrum according to some *scoring function.* This requires us to discuss the meaning of a scoring function. Basically, for a given real experimental spectrum, we can find some candidate peptides for this spectrum. We will discuss how to find the candidate peptides in later sections. Here, we will assume that the candidate peptides are obtained. For each candidate peptide, a hypothetical spectrum can be generated *in silico*, based on the assumption of some fragmentation patterns and frequencies. Different fragmentation patterns will generate different ion types. The usual ion types are b-ions and y-ions, as shown in Figure 1. The **scoring function** then measures the similarity of the real spectrum to the hypothetical spectrum of each candidate peptide. The candidate peptide associated with the hypothetical spectrum that shows the most similarity with the real spectrum is then reported as the best candidate peptide that explains the real spectrum. Sometimes the similarity score is associated with a p-value, which gives a probability that the score is achieved by random chance. The design of a good scoring function is never an easy task and is an active research area in which considerable efforts have been made to find a good scoring function [7-14].

To illustrate the abstract description of the scoring function above, we will give a brief description of one of the scoring methods, a four-step process called the SEQUEST algorithm [7]. The algorithm begins (step 1) with tandem mass spectrometry data reduction. In this

step, fractional mass-to-charge ratios are rounded to the nearest integers, and then a 10-u window around the precursor ion is removed to avoid matching to the unfragmented precursor ion. The 200 most abundant ions are selected for scoring purposes. In step 2, the candidate peptide sequences are then identified by matching the precursor peptide masses to the masses of all possible peptides in a protein database. Peptides with ± 3 u or ± 1 u are selected as candidate peptides. One hypothetical spectrum is then generated for each candidate peptide, and these hypothetical spectrums are compared with the real spectrum to produced a preliminary ranked list of 500 best-fit sequences (step 3). This preliminary ranked list considers the number of matching ions within the mass tolerance of ± 1 u, the abundances of the matching ions, the continuity of an ion series, and the presence of immonium ions for the amino acids His, Tyr, Trp, Met, and Phe. Finally, in step 4, these 500 sequences are then subject to a cross-correlation analysis to generate the final ranked list.

Rationales for *de novo* sequencing

Now, let us return to the rationales for doing *de novo* peptide sequencing. Usually, searching against a sequence database is the first choice for peptide identification, because the candidate peptides can be found from the database. However, *de novo* peptide sequencing comes into play in various situations. First, the protein of interest might not be present in the sequence database. One case is that the sequence database is incomplete, which is the situation for many model animals and plants. Another case is that the protein of interest might be a novel protein, which might appear when pharmaceutical scientists try to screen for new drugs from synthetic

proteins. Second, there are prediction errors in gene-finding programs. Thus, it might not be possible to find the true protein from the predicted protein database. Third, some scientists might want to study the proteome before the genome, in which case no sequence database might be available. In a fourth situation, genes might undergo alternative splicing, which would result in novel proteins. The occurrence of single nucleotide polymorphisms (SNPs) in coding regions may also lead to different protein variants. In a fifth case, *de novo* sequencing can be helpful for studying amino acid mutations and protein modifications. Finally, when a database search generates ambiguous results, *de novo* sequencing can be used as a validation tool.

Early development of *de novo* sequencing algorithms

Over the years, various algorithms have been developed to address the *de novo* sequencing problem. One naive method [15, 16], is to list all possible candidate peptides according to the mass of the parent ion of the tandem mass spectrum. This is sometimes called *exhaustive listing*. All of the candidate peptides are then compared with the real spectrum to find out which one is the best match. One computational difficulty inherent in this approach is that there will most likely be a large number of possible candidates for a typical parent mass that may range from below one thousand to several thousand Daltons [15, 16]. For example, as described in Reference 15, for a parent peptide of molecular weight of 774, there will be 21,909,046 possible candidate peptides.

An alternative approach, sometimes called "subsequencing," has also been tried on mass

spectra data generated by various mass spectrometers [17-20]. In this approach, short sequences that represent only a portion of the whole sequence are tested against the real spectrum. Those subsequences that account for some observed ions are then extended one residue at a time until the whole sequence is tested. During the subsequence extension, only those subsequences that have significant matching with the real spectrum are retained. One disadvantage of this approach is that some good candidate peptides might be discarded when some regions of a peptide are less represented by fragmentation ions. It is important to be aware that the fragmentation frequencies of a peptide are usually not evenly distributed over the whole peptide.

A third method employs the use of graphical display of the data [21]. In this method, fragmentation ions that differ by the mass of one amino acid are represented by connected lines, thus allowing the visualization of ion series of the same type. Such an approach is not quite suitable for high-throughput environments, but this method can be quite helpful for manual *de novo* interpretation of tandem mass spectra.

A fourth approach uses graph theory [22-25, 13]. This approach has been proven to be quite successful and will be discussed in the next session.

Application of graph theory in mass spectrometry

An application of graph theory in *de novo* peptide sequencing was first proposed by Bartels [22]. The basic idea is to transform a real spectrum into a graph called "*Spectrum Graph*". For the transformation, each peak in the real spectrum is represented as a vertex (or several

vertices) in the spectrum graph, and a directed edge is established between two vertices if the mass difference of the two vertices equals the mass of some amino acids. Various algorithms have been designed to find paths in the spectrum graph in which the corresponding peptides provide a good explanation of the real spectrum. Some algorithms of this type are introduced below

The Lufefisk algorithm. The Lutefisk algorithm was designed by Taylor and Johnson [25]. In this algorithm, the authors first reduced the real spectrum data to a list of significant fragment ions. Then they determined the N- and C-terminal evidence lists, which provide the evidence for the possible N-terminal and C-terminal ions, respectively. After the N-terminal and C-terminal evidence lists were obtained, a "sequence spectrum" was formed, in which the x ordinate consisted of the m/z values for the b-ions and the y ordinate consisted of the probability of cleavage at each site. The program then proceeded by tracing out sequences, starting from the N-terminus, by finding b-ion values that differed from the N-terminus by the mass of one or several amino acids. After all the sequences had been obtained, a scoring procedure was carried out to rank the sequences.

The SHERENGA algorithm. The SHERENGA algorithm was developed by Dancik *et al.* [13]. Because the creation of a spectrum graph is based on the ion types, the authors designed a method to automatically learn ion types from a training set of experimental spectra of known sequences, without knowing *a priori* the fragmentation patterns. After the ion types were learned, they transformed the real spectrum into a spectrum graph using the following steps. First, the ion types were represented by $\triangle = \{\delta_1, ..., \delta_k\}$, where k is the number of ion types and each δ_i represents the offset of the corresponding ion type. The real spectrum S was

then transformed into a spectrum graph $G_{\Delta}(S)$ as follows. Each peak s of the real spectrum S generated k vertices $V(s) = \{s+\delta_1, ..., s+\delta_k\}$. Two vertices, u and v, were then connected by a directed edge from u to v if $v - u$ equaled the mass of some amino acid. The peptide sequencing problem was then cast as the *longest path problem in a directed acyclic graph.* The authors also pointed out that the longest path may correspond to unrealistic solutions because it may use multiple vertices associated with the same real spectral peak. One solution is to find the longest anti-symmetric path. They claimed there exists an efficient algorithm to find the anti-symmetric longest path in the spectrum graph.

Dynamic programming and suboptimal concept

As pointed out by Dancik *et al.* [13], the problem of finding the longest path in a directed acyclic spectrum graph while avoiding multiple assignments to the same peak is NP-complete in the general case [26]. However, Dancik *et al.* [13] and Chen *et al.* [27] observed that there is a *special structure* for *forbidden pairs* of vertices (twins) in the spectrum graph. That is, the forbidden pairs are *non-interleaving.* Two forbidden pairs of vertices (x_1, y_1) and (x_2, y_2) are *non-interleaving* if the intervals (x_1, y_1) and (x_2, y_2) do not interleave. Chen *et al.* [27] then proposed a dynamic programming approach to find the anti-symmetric longest path in the spectrum graph.

Dynamic programming is a common technique in solving optimization problems [26]. In dynamic programming, an optimization problem is solved in a bottom-up fashion, through combining the solutions to sub-problems, which gives an optimal solution for the original

problem. Chen *et al.* [27], employing dynamic programming, provided a polynomial time algorithm for the *de novo* sequencing problem using a spectrum graph. Following is an outline for the algorithm. First, an NC-spectrum graph G ("NC" denotes "N-terminal and C-terminal") is constructed from the real spectrum, by assuming that each peak in the real spectrum can be a b-ion or a y-ion. Thus, each peak in the real spectrum will generate two vertices in the NC-spectrum graph G. All of the vertices are then placed on the real line at positions corresponding to the mass values of the vertices. If the mass difference between two vertices u and v equals the total mass of some amino acid residues, a directed edge is drawn between u and v, pointing from the low-mass vertex to the high-mass vertex. To give the reader a feel for how an NC-spectrum graph can be generated from a mass spectrum, an example is given in Figure 2. Next, the nodes of G are renamed in an order from left to right as $x_0, x_1, \ldots, x_k, y_k, \ldots, y_1, y_0$, where every pair, x_i and y_i, $1 \leq i \leq k$, corresponds to two mutually exclusive assumptions of the same mass peak. The dynamic programming algorithm then finds a path with the maximum path score from x_0 to y_0 that contains the edge (x_i, y_j), $i \neq j$.

The dynamic programming algorithm will find the optimal solution. However, the optimal solution may not be the real sequence that produces the real spectrum. Even database search programs sometimes report several sequences with similar scores because the scoring function can misinterpret the spectral data. Noise and unknown ions may also be interpreted as real ions by the programs. For these reasons, the suboptimal solutions are of great interest because they might give us the real sequence that generated the spectrum. Lu and Chen [28] further explored this application of suboptimal solutions in the dynamic programming

algorithm. In this suboptimal algorithm, a real spectrum with k peaks is transformed into a *matrix spectrum graph* $G = (V, E)$, where $|V| = O(k^2)$ and $|E| = O(k^3)$. A polynomial time suboptimal algorithm was then proposed to find all of the suboptimal solutions (peptides) in $O(p|E|)$ time, where p is the number of solutions.

Inspired by the work of Dancik *et al.* [13] and Chen [27], two more research groups also proposed dynamic algorithms to solve the *de novo* peptide sequencing problem [29, 30]. The work by Bafna and Edwards [29] also considered the suboptimal solutions in their dynamic programming machinery.

For interested readers, a list of *de novo* peptide sequencing programs is listed in Table 1.

Concluding remarks

Mass spectrometry has become an important tool for proteomics study. Normally, searching against a sequence database is the first choice for protein identification. However, *de novo* sequencing comes into play in various situations. Over the years, numerous computer algorithms have been developed for *de novo* peptide sequencing, and there are also manual methods for interpretation of mass spectra. For example, one strategy for the interpretation of mass spectra generated from tryptic peptides was presented by Kinter and Sherman [31].

At the same time, the *de novo* sequencing problem via tandem mass spectrometry is still not solved in general. In other words, the information contained in tandem mass spectrometry cannot be readily converted into a fully unambiguous peptide sequence. Usually, a *de novo* sequencing program has been designed for some machine-dependent tandem mass spectra and

thus is not universally applicable to spectra generated by other types of mass spectrometers. To our knowledge, none of the current *de novo* sequencing programs take into account internal fragmentation of the parent ion. Also, current *de novo* sequencing programs usually assume that all daughter ions have a +1 charge state, which might not be the real case.

Nevertheless, the development of computational algorithms, including *de novo* peptide sequencing methods, database search algorithms, and other computational tools, together with mass spectrometric instrumental developments, would substantially enhance our capabilities in biological studies.

References

1. Link, A.J. *et al.* (1999) Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology.* **17**: 676-682

2. Washburn, M.P. *et al.* (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology.* **19**: 242-247

3. Gavin A. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature.* **415**: 141-147

4. Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature.* **415**: 180-183

5. Petricoin, E.F. *et al.* (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet.* **359**:572-577

6. Peng, J. *et al.* (2003) Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *Journal of Proteome Research.* **2**: 43-50

7. Eng, J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of American Society for Mass Spectrometry.* **5**: 976-989

8. Mann, M. and Wilm, M. (1994) Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry.* **66**: 4390-4399

9. Clauser, K.R. *et al.* (1999) Role of accurate mass measurement (+/- 10ppm) in protein identification strategies employing MS or MS/MS. *Analytical Chemistry.* **71**: 2871-2882

10. Perkins, D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis.* **20**: 3551-3567

11. Bafna, V. and Edwards, N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics.* **17** Suppl 1: S13-21

12. Field, H.I. *et al.* (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics.* **2**: 36-47

13. Dancik, V. *et al.* (1999) *De novo* peptide sequencing via tandem mass spectrometry: a graph-theoretical approach. *Journal of Computational Biology.* **6**: 327-342

14. Havilio, M. *et al.* (2003) Intensity-based statistical scorer for tandem mass spectrometry. *Analytical Chemistry.* **75**: 435-444

15. Sakurai, T. *et al.* (1984) PAAS 3, a computer program to determine probable sequence of peptides from mass spectrometric data. *Biomedical & Mass Spectrometry.* **11**: 396-399

16. Hamm, C.W. *et al.* (1986) Peptide sequencing program. *Computer Applications in the Biosciences.* **2**: 115-118

17. Bieman, K. *et al.* (1966) Determination of the amino acid sequence in oligopeptides by computer interpretation of their high-resolution mass spectra. *Journal of American Chemical Society.* **88**: 5598-5606

18. Ishikawa, K. and Niwa, Y. (1986) Computer-aided peptide sequencing by fast-atom-bombardment mass-spectrometry. *Biomedical & Environmental Mass Spectrometry.* **13**: 373-380

19. Siegel, M.M. and Bauman, N. (1988) An efficient algorithm for sequencing peptides using fast atom bombardment mass-spectral data. *Biomedical & Environmental Mass Spectrometry.* **15**: 333-343

20. Johnson, R.S. and Biemann, K. (1989) Computer-program (seqpep) to aid in the interpretation of high-energy collision tandem mass-spectra of peptides. *Biomedical & Environmental Mass Spectrometry.* **18**: 945-957

21. Scoble, H.A. *et al.* (1987) A graphics display-oriented strategy for the amino-acid sequencing of peptides by tandem mass-spectrometry. *Fresenius Zeitschrift Fur Analytische Chemie.* **327**: 239-245

22. Bartels, C. (1990) Fast algorithm for peptide sequencing by mass spectrometry. *Biomedical & Environmental Mass Spectrometry* **19**: 363-368

23. Hines, W.M. *et al.* (1992). Pattern-based algorithm for peptide sequencing from tandem high-energy collision-induced dissociation mass-spectra. *Journal of the American Society for Mass Spectrometry.* **3**: 326-336

24. Fernandez-de-Cossio, J. *et al.* (1995) A computer program to aid the sequencing of peptides in collision- activated decomposition experiments. *Computer Applications in the Biosciences.* **11**: 427-434

25. Taylor, J.A. and Johnson, R.S. (1997) Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry.* **11**: 1067-1075

26. Cormen, T.H. *et al.* (2001) *Introduction to Algorithms*, The MIT Press

27. Chen, T. *et al.* (2001) A dynamic programming approach for *de novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology.* **8**: 325-337

28. Lu, B. and Chen, T. (2003) A suboptimal algorithm for *de novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology.* **10**: 1-12

29. Bafna, V. and Edwards, N. (2003) On *de novo* interpretation of tandem mass spectra for peptide identification. Annual Conference on Research in Computational Molecular Biology (RECOMB'03) 1-8

30. Ma, B. *et al.* (2003) An effective algorithm for the peptide *de novo* sequencing from MS/MS spectrum. *CPM 2003*: 266-277

31. Kinter,M. and Sherman,N.E. (2000) *Protein sequencing and identification using tandem mass spectrometry*, John, Wiley & Sons, Inc.

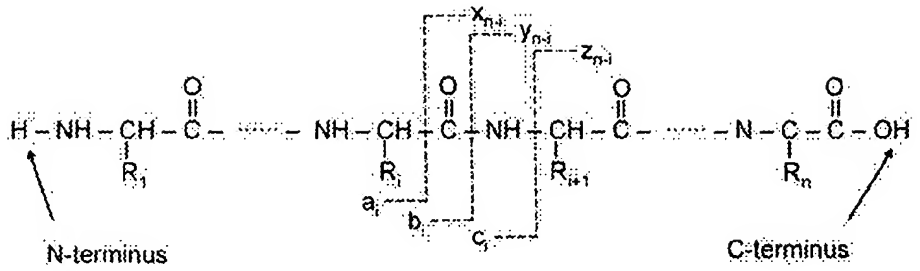


Figure 1: A peptide of n amino acids and possible fragmentation patterns. When the peptide is fragmented between the i-th and $i+1$-th amino acids, three possible fragmentations can happen, yielding three pairs of ions $(a_i,\ x_{n-i})$, $(b_i,\ y_{n-i})$, and $(c_i,\ z_{n-i})$, respectively. The a-ions, b-ions and c-ions are N-terminal ions while x-ions, y-ions and z-ions are C-terminal ions. The most common ion types are b-ions and y-ions.

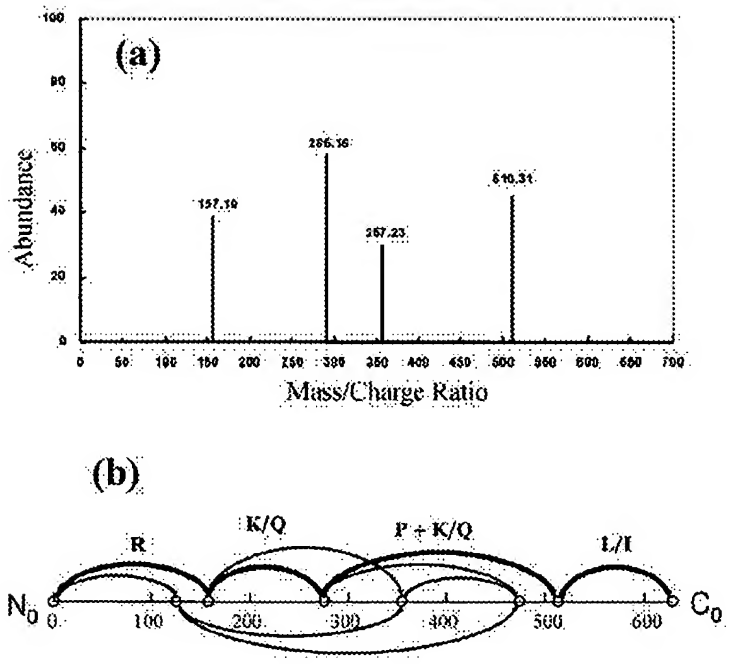


Figure 2: An NC-spectrum graph example. (a) An artificial tandem mass spectrum of the peptide NDEMK(617.25 Daltons). (b) A sparse NC-spectrum graph constructed from the spectrum shown in A. One of the paths running from N_0 to C_0 is bolded and labelled with possible corresponding amino acids. The symbol "+" means "and", while "/" means "or". For example, "E+M" means "E and M (order undetermined)", while "Q/K" means "Q or K".

Table 1: A list of *de novo* peptide sequencing programs.

	URL	Reference
1	http://www.protein.osaka-u.ac.jp/rcsfp/profiling/Seqms/SeqMS.html	[24]
2	http://hto-c.usc.edu:8000/msms/	[28]
3	http://www.bioinformaticssolutions.com/software/peaks/	[30]